NUI Galway
OÉ Gaillimh

ARAN - Access to Research at NUI Galway

Provided by the author(s) and NUI Galway in accordance with publisher policies. Please cite the published version when available.

Title	Drug target discovery using knowledge graph embeddings
Author(s)	Mohamed, Sameh K.; Nováek, Vít; Nounu, Aayah
Publication Date	2019-04-08
Publication Information	Mohamed, Sameh K., Nováek, Vít, & Nounu, Aayah. (2019). Drug target discovery using knowledge graph embeddings. Paper presented at the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), Limassol, Cyprus, 08-12 April.
Publisher	Association for Computing Machinery
Link to publisher's version	https://doi.org/10. 1145/3297280.3297282
Item record	http://hdl.handle.net/10379/15065
DOI	http://dx.doi.org/10. 1145/3297280.3297282

Downloaded 2019-09-16T03:13:24Z

Some rights reserved. For more information, please see the item record link above.

Drug Target Discovery Using Knowledge Graph Embeddings

Sameh K. Mohamed
Data Science Institue
Insight Centre for Data Analytics
National University of Ireland Galway
sameh.kamal@insight-centre.org

Aayah Nounu
MRC Integrative Epidemiology Unit
University of Bristol
An0435@bristol.ac.uk

Vít Nováček
Data Science Institue
Insight Centre for Data Analytics
National University of Ireland Galway
vit.novacek@insight-centre.org

ABSTRACT

The field of drug discovery has entered a plateau stage lately. It is increasingly more expensive and time-demanding to introduce new drugs into the market. One of the main reasons is the slow progress in finding novel targets for drug candidates and the lack of insight in terms of the associated mechanisms of action. Current works in this area mainly utilise different chemical, genetic and proteomic methods, which are limited in terms of the scalability of experimentation and the scope of studied drugs and targets per experiment. This is mainly due to their dependency on laboratory experiments and available physical resource. This has led to an increasing importance of computational methods for the identification of candidate drug targets. In this work, we introduce a novel computational approach for predicting drug target proteins. We approach the problem as a link prediction task on knowledge graphs. We process drug and target information as a knowledge graph of interconnected drugs, proteins, disease, pathways and other relevant entities. We then apply knowledge graph embedding (KGE) models over this data to enable scoring drug-target associations, where we employ a customised version of state-of-the-art KGE model ComplEx. We generate a benchmarking dataset based on KEGG database to train and evaluate our method. Our experiments show that our method achieves best results in comparison to other traditional KGE models. Specifically, the method predicts drug target links with mean reciprocal rank (MRR) of 0.78 and Hits@10 of 0.88. This provides a promising basis for further experimentation and comparisons with domain-specific predictive models.

CCS CONCEPTS

• **Semantic networks**; • **Machine Learning**; • **Machine learned ranking**;

KEYWORDS

Drug Target Discovery, Knowledge Graph Embeddings, Link Prediction

ACM Reference Format:
Sameh K. Mohamed, Aayah Nounu, and Vít Nováček. 2019. Drug Target Discovery Using Knowledge Graph Embeddings. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3297280.3297282

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC '19, April 8–12, 2019, Limassol, Cyprus
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297282

1 INTRODUCTION

The development of drugs has a long history [6]. Until quite recently, pharmacological effects were often discovered using primitive trial and error procedure, applying plant extracts on living system and observing the outcomes. Later, drug development evolved to elucidating mechanisms-of-actions of drug substances and their effects on phenotype. The ability to pharmacologically isolate active substances was a key step towards modern drug discovery [34, 35]. More recently, advances in molecular biology and biochemistry allowed for more complex analysis of drugs, their targets and their mechanisms of action. The study of drug targets has become very popular, where studies utilise different chemical genetic [35] and proteomic methods [33] such as affinity chromatography and expression cloning approaches. These, however, can only process limited number of possible drugs and targets due to dependency on laboratory experiments and available physical resource. Computational approaches have therefore been extensively studied lately [17, 18, 42].

In this work we introduce a specific computational approach for predicting drug target proteins. Our objective is to score possible associations between drugs and proteins according to the probability of the association holding true. The ultimate goal is to assist lab experimentation in narrowing the scope of possible new drug targets investigated. In the current drug target knowledge bases like DrugBank [40] and KEGG [12], information about drugs contains their relationship with target proteins (or their genes), action pathways and targeted diseases. These components are represented as graphs form of interconnected entities and relations. Such data can naturally be interpreted as a knowledge graph, where the task of finding new associations between drugs and their targets can be formulated as a link prediction task. In this context, knowledge graph embedding (KGE) models are a fit natural application, where they are known to provide state-of-the-art results in link prediction on knowledge graphs [23]. Despite the growing body of computer simulation based drug target prediction frameworks [39, 42, 43], none of these works utilise knowledge graph embedding models in their predictive pipelines.

The objective of this work is to demonstrate the usefulness of knowledge graph embedding models in the area of drug target prediction. We also identify KGE techniques that can provide the best accuracy in predicting drug targets. This is presented as a stepping stone towards a domain-specific KGE-based drug target prediction model and its extensive comparison with existing related models such as the DDR [25] and the DNILMF [11] models.

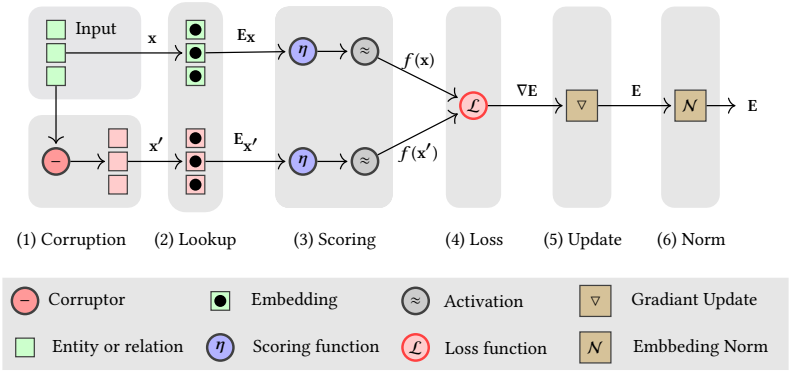

Figure 1: Phases of one epoch training of a KGE model over one training instance

Our knowledge graph embedding approach, ComplEx-SE (Complex embeddings with squared error loss), is a customisation of the state-of-the-art KGE model ComplEx [38] with a square error-based training objective. We build a drug-target centred dataset from KEGG database to train and evaluate our model, and we show by experiments that it achieves best results in terms of predicting new drug target links. To the best of our knowledge, there are currently no models for discovering new drug targets using KGE model for link prediction on biomedical knowledge graph, so we evaluate our method compared to other state-of-the-art KGE models like the Translating Embeddings (TransE) model [2], the DistMult model [44] and the Complex Embedding (ComplEx) model [38].
The rest of this paper is structured as follows: section 2 discuss the problem of drug target discovery and presents fundamental background concepts about KGE models and their evaluation metrics. Section 4 present the subset of KEGG that we use, and discusses its component. Section 5 presents our KGE approach and its base work the ComplEx model. Section 6 present the experimental setup, the evaluation protocol, section 3 discusses similar related works that uses computational-based approaches for finding drug targets and section 7 discusses the results of our experiments and lesson learnt. We finally present our conclusions and possible future works in section 9.

2 BACKGROUND

In this section, we discuss the advantages and implications of finding drug targets that are not yet known. We also discuss modelling information in knowledge graphs, the underlying concepts of knowledge graph embedding models and their evaluation techniques.

2.1 Drug Target Discovery

The process of discovering and developing drugs with one gene target requires time and money. Rarely does a drug only bind to its intended target, but rather off-target effects are common [41], and this may lead to unwanted side-effects [3]. Conversely, the off-target effects may be useful for drug-repurposing reasons. Drug repurposing is defined as the use of approved drugs for new diseases [4]. It is believed to take around 10-17 years from the conception of a drug to when it becomes a licensed treatment for disease, with a success rate of less than 10% [1]. Drug repurposing is advantageous as the safety profile of the drug is already known and reduces the time and cost required to bring a new drug into the clinic [4].

The identification of new protein targets also allows the development of drugs that specifically target the protein of interest. For example, *aspirin* is currently being considered for use as a chemopreventative agent [27, 30, 31] but there are concerns with regards to side-effects caused by its long-term use, such as upper gastrointestinal bleeding [16]. By identifying the exact protein targets of *aspirin*, new drugs can be developed specifically for these proteins to avoid the unwanted side-effects.

The use of computational approaches is useful as they are free from bias and are therefore not influenced by prior knowledge and opinions, unlike laboratory-based methods. These approaches bypasses the need to spend a long amount of time in the laboratory and can be used to provide guidance on the direction of research within the laboratory, therefore saving both time and money. Follow-up experiments can then be carried out in the laboratory for confirmation of the new proteins targeted by the drug, allowing direct conclusions to be made of treatment effect.

Overall, computational approaches are considered useful methods to identify off-target interactions and can also be used for the possibility of drug repurposing. As they reduce the time required to manually discover other unintended protein targets and may reduce the large costs required in doing so.

2.2 Knowledge Graphs

Knowledge graphs are a data representation that model relational information as a graph, where the graph nodes represent knowledge entities and its edges represent relations between them. They model facts as (subject, predicate, object) (SPO) triples *e.g.* (*Aspirin*, *Drug-Target*, *COX-1*), where a subject entity is connected to an object entity through a predicate relation.

In recent years, knowledge graphs have become a popular means for data representation in the semantic web community to create the "web of data", which is a network of interconnected entities

that can be easily interpreted by both humans and machines [36], where knowledge graphs are used to model linked data. They have also been used as convenient means for modelling information in many different domains, including general human knowledge [15], biomedical information [7] and language lexical information [19]. Knowledge graphs are now used in different applications such as enhancing semantics of search engine results [26, 32], biomedical discoveries [21], or powering question answering and decision support systems [8].

2.3 Knowledge Graph Embedding

Knowledge graph embedding models learn a low rank vector representation of knowledge entities and relations that can be used to rank knowledge assertions according to their factuality. KGE models are trained in a multi-phase procedure as shown in Fig. 1, where their objective is to effectively learn a vector representation of entities and relations that can be used to score and rank possible knowledge facts.

First, a KGE model initialises all embedding vectors using random noise values. It then uses these embeddings to score the set of true and false training facts using a model-dependent scoring function. The output scores are then passed to the training loss function to compute training error as shown in Fig. 1. These errors are used by optimisers like AMSGrad [28] to generate gradients and update the initial embeddings, where the updated embeddings give higher scores for true facts and less for false facts. This procedure is performed iteratively for a set of iterations *i.e.* epochs in order to reach a state where embeddings provide best possible scoring for both true and false possible facts.

2.4 Ranking Metrics

In the following, we present the metrics that we use in the evaluation of our approach.

(1) Mean reciprocal rank (MRR): This is the harmonic mean of the rank position of the first relevant element, and it is defined as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

where $rank_i$ refers to the rank position of the first relevant element for the i-th query. The output values of mean reciprocal rank are normalised from 0 to 1, where 1 represents perfect ranking and decaying values towards 0 represent decreasing accuracy.

(2) Hits@k: This is the number of correct elements predicted among the top-k elements, where we use Hits@1, Hits@3 and Hits@10. This metric indicates that the model's probability of ranking the relevant (true) fact in the top k element scores in the rank.

3 RELATED WORK

In this section we discuss related works where we target two kinds of activities. Firstly, other computer based approaches for predicting drug targets. Secondly, relation link prediction approaches and state-of-the-art knowledge graph embedding models.

3.1 Computer Based Drug Target Prediction

Yamanishi et al. [42] developed one of the early computational approaches to predict drug targets, where their approach utilised a statistical model that infers drug targets based on a bipartite graph of both chemical and genomic information. More recent works like COSINE [29] and NRLMF [17] approaches introduced the use of drug-drug and target-target similarity measures to infer possible drug targets. These approaches enabled new drugs and drug targets with limited or no information about their interaction data since they depend on the drug-drug and target-target similarities. However, these methods only utilised a single measure to model components similarity.

Other drug target prediction models like KronRLS-MKL [22] and BLM-NII [18] integrated different similarity measures to model the similarity between drugs and targets. These approaches use both linear and non-linear combinations of similarity measures to encode the similarity between drugs and their targets, where non-linear combinations provided better predicting drug-target predictions [18].

Recently, Hao et al. [11] proposed a model that uses matrix factorisation to predict drug targets over drug information networks. Their model, DNILMF, operates in a four-step procedure. First, it infers different profiles for both drugs and targets and constructs kernel matrices for these profiles. It then diffuses drug profiles kernel matrices with their structure kernel matrices. It then diffuses target profiles kernel matrices with their sequence kernel matrices. Finally, the DNILFM model uses the outputs of the previous steps to predict drug targets based on their network neighbours. This approach showed significant predictive accuracy improvements over other methods on standard benchmarking datasets [11, 42].

In most recent times, the current state-of-the-art work on computation drug target discovery is the DDR model [25], which predicts drug targets using heterogeneous graphs that contain drug target interactions in a multiple phases procedure. First, it computes similarity indices for drugs and their targets. It then selects a subset of these similarities in a heuristic process to obtain optimal combinations of similarities. Finally, it combines selected similarities using a non-linear fusion technique, and combines diffusion output with random walk features from the heterogeneous graphs to predict drug targets. Despite the complexity of the the DDR model, it currently provides state-of-the-art results in predicting drug targets using computational approaches [25].

3.2 Link Prediction in Knowledge Graph

In recent years, various predictive frameworks were developed to predict new links in knowledge graphs, where these frameworks serve in various applications such as semantic search engines [26, 32], biomedical discoveries [21], and question answering systems [8]. Link prediction models can be categorised into two categories: graph-feature based models and latent-feature based mode. Graph-feature based models utilise graph features like paths and graph patterns to predict possible connecting links between graph entities. For example, the path ranking algorithm (PRA) [14] uses

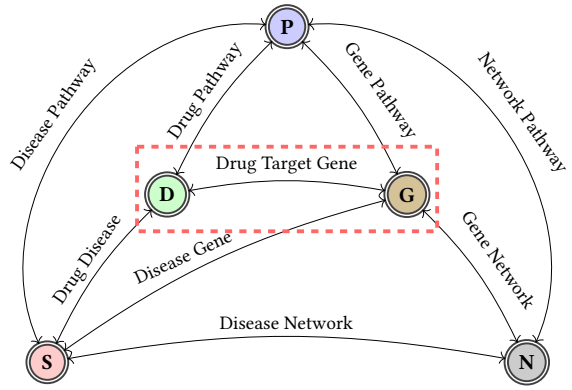

Figure 2: Knowledge graph about drugs, their target genes, pathways, diseases and gene variant networks extracted from KEGG.

connecting paths between entities generated by random walks to infer possible links between them, where other models like the subgraph feature extraction model (SFE) [9] and the distinct subgraph path (DSP) [20] employ a combination of connecting path and subgraph paths of two entities to predict their possible associations. On the other hand, latent-feature based models *i.e.* knowledge graph embedding models, use a generative approach to learn low-rank embeddings for knowledge entities and relations in order to score their possible associations. These approaches use multiple techniques like tensor factorisation as in the DistMult model [2] and latent distance similarity as in the TransE model [44] to model possible interactions between graph embeddings and provide scores for possible graph links. For further information on both approaches, Nickel et al. [23] provides an extended review for both graph-feature based and latent-feature based models in the task of link prediction in knowledge graphs.

4 DATA

In this section we discuss the KEGG database [12, 13] with focus on the components that we use to train and evaluate our approach.

KEGG is a knowledge base that contains information about biological systems like cells and organisms at the molecular level. It contains different types of biological information entities like genes, pathways, drugs, disease, etc. The data in KEGG is structured as a network of inter-connected entities that resembles the biological eco-system at the molecular level. In our study, we focus on information related drugs and their targets, where we only the following KEGG components are considered:

(1) Drugs: The KEGG drug database[1] is a comprehensive drug information resource for approved drugs. It contains multiple types of information about drugs such as the chemical structure, associated targets, action pathways and targeted diseases. In our study, we only consider drug associations to the elements specified in Table 1.

Table 1: Statistics of objects and their inter-connections in the subset of KEGG dataset that we use.

Object	Count	Drug	Gene	Pathway	Disease	Network
Drug	4670	•	12004	7910	2160	0
Gene	8881	12004	•	497	239	4534
Pathway	329	7910	497	•	1803	524
Disease	1873	2160	239	1803	•	441
Network	448	0	4534	524	441	•

(2) Genes: The KEGG Gene database[2] contains information about genes, their sequences and their associations with other biological entities. While drug targets in the living systems are usually proteins, the KEGG database uses genes as a representation of drug targets, where genes represent their product proteins. In the rest of this study, we use the KEGG genes to represent product proteins as drug targets in the knowledge graph we have created.

(3) Pathways: The KEGG database also contains information about biological pathways associated with manually curated maps of their reactions. The pathway database[3] in KEGG includes pathways of different activities such metabolism, environmental information processing, human disease, etc. Each pathway is associated to its related entities *e.g.* genes, drugs and diseases, where we use such associations to construct our knowledge about pathways.

(4) Diseases: The KEGG disease database[4] is structured in a similar form as in the previously mentioned entity databases, where our main interest is to utilise the associations between disease and our other investigated entities. However, the disease database contains associations to other entity types such as *carcinogens* that can be helpful to extend the knowledge about specific cancerous disease in future studies.

(5) Networks: The KEGG network database [5] contains information on the perturbation of human genes, where it encodes knowledge about the different variants and other perturbants of human genes that are involved in the perturbed molecular reaction networks. Similarly, instances of the network database are linked to their related entities in other KEGG databases.

In our study, we gather all the possible associations between the previously mentioned KEGG entities to generate a biological knowledge graph that is centred around drugs and their target genes as shown in Fig. 2. The statistics of the counts of each entity type and the inter-connecting links between entities are provided in Table 1.

[1] https://www.genome.jp/kegg/drug/

[2] https://www.genome.jp/kegg/genes.html
[3] https://www.genome.jp/kegg/pathway.html
[4] https://www.genome.jp/kegg/disease/
[5] https://www.kegg.jp/kegg/network.html

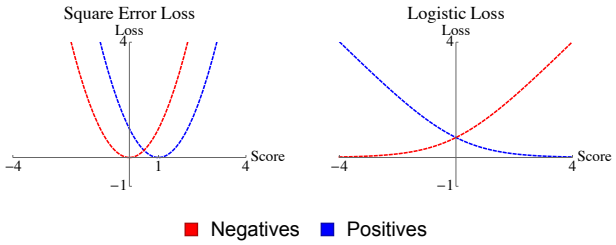

Figure 3: Plots of loss growth of squared error and logistic pointwise loss functions compared to scores of positive and negative instances.

5 OUR APPROACH

In this section, we present the technical details of our approach, which is a modification upon the the ComplEx knowledge graph embedding model [38]. We discuss the ComplEx model, its scoring and loss functions and our modifications.

5.1 Complex Scoring Function

The ComplEx model is a tensor factorisation based knowledge graph embedding model. It represents knowledge entities and relation using complex vector embeddings, where each embedding is represented by two vectors (real and imaginary). Fact assertions in the ComplEx model are evaluated using a factorisation based scoring function that is defined as follows:

$$f_{\text{ComplEx}}(s,p,o) = \sum_{k=1}^{K} \text{Re}(< e_{sk}, e_{rk}, \overline{e_{ok}} >)$$

where e_s, e_r and e_o are the embeddings of the subject, the relation and the object respectively, e_{sk} is the k-th component of the embedding e_s, K is the embedding size (vector length), $Re(x)$ is the real part of complex value x and $\overline{x}$ is the complex conjugate of x such that $\overline{x} = a - ib$ if $x = a + ib$. This formulation can be further relaxed as follow:

$$f_{\text{ComplEx}}(s,p,o) = \sum_{k=1}^{K} e^r_{sk} e^r_{rk} e^r_{ok} + e^i_{sk} e^r_{rk} e^i_{ok} + e^r_{sk} e^i_{rk} e^i_{ok} - e^i_{sk} e^i_{rk} e^r_{ok}$$

where x^r and x^i are the real and imaginary parts of complex value x respectively. The use of the complex conjugate in the ComplEx model allows it to encode embedding interactions in an asymmetric operation, which enables it to model facts with both symmetric and asymmetric predicates unlike other factorisation based models like the DistMult model [44].

5.2 Training Objective

In the task of link prediction, knowledge graph embedding models are considered learning to rank models, where they employ traditional ranking functions *e.g.* pointwise and pairwise ranking losses to model their training loss. The ComplEx model by default uses a pointwise ranking loss with a negative-logistic transformation to model its training loss, which is defined as follows:

$$\underset{\text{logistic}_{Pt}}{\mathcal{L}} = \sum_{x \in T} \log(1 + \exp(-l(x) \cdot f(x))), \tag{1}$$

where x is an (s, p, o) fact and T is the set of all training facts with negative samples and $l(x)$ is the true label of fact x such that x is equal to 1 when true and -1 otherwise. This allows the models to effectively update the embeddings of both entities and relations to give high scores to true facts and lower scores to false facts as shown in Fig. 3.

5.3 New Loss Objective for The ComplEx model

Trouillon and Nickel [37] have shown that the choice of objective loss in KGE models has a huge impact on their predictive accuracy. They showed that despite the equivalence of both the ComplEx and the Holographic embedding (HolE) [24] models, they vary in accuracy due to their dependency on different training loss objectives. This difference is caused by the fact that the HolE model uses a max-margin loss while the ComplEx model uses a log-likelihood loss. Following their remarks, in this work, we propose a new loss objective to the ComplEx model, and we show that it suit the limited size and number of predicates in our dataset. We propose a new loss function based on the square error of the difference between ComplEx scores assertions and their true labels using a 0 and 1 labelling such that 0 represents false fact assertions and 1 represents true fact assertions. The new square error based loss is defined as follows:

$$\underset{\text{SE}_{Pt}}{\mathcal{L}} = \sum_{x \in T} \frac{1}{2}(l(x) - f(x))^2 \tag{2}$$

where $l(x)$ is the label of fact x with $l(x) = 0$ if x is false and 1 otherwise. This also allows the square error loss to force embedding updates that produce normalised scores around 0 and 1 unlike the logistic loss with an open range of scores.

We prove that our new representation of ComplEx model training loss outperforms its default version using an empirical evaluation framework described in Section 6. In the following, we discuss some properties of both the logistic and square error based losses.

Fig. 3 shows the differences between the growth of both the square error based loss and ComplEx's default logistic loss, where both functions show different loss growth rates which affect the growth rate of the values of their output gradients. The logistic loss has a linear growth rate and its gradient per single instance is defined as follows:

$$\Delta_x = \exp(f(x))/[1 + \exp(f(x))] \tag{3}$$

where this form grows in a sub-linear sigmoid fashion. This limits the output gradients for each training instance in the range of $[0, 1]$. On the other hand, the square growth of square error loss yields linearly growing gradients defined as follows:

$$\Delta_x = \begin{cases} 0 & \text{for } f(x) - l(x) = 0 \\ f(x) & \text{for } f(x) - l(x) < 0 \\ -f(x) & \text{for } f(x) - l(x) > 0 \end{cases} \tag{4}$$

which enables the ComplEx model to produce gradients within the range $[0, \infty)$.

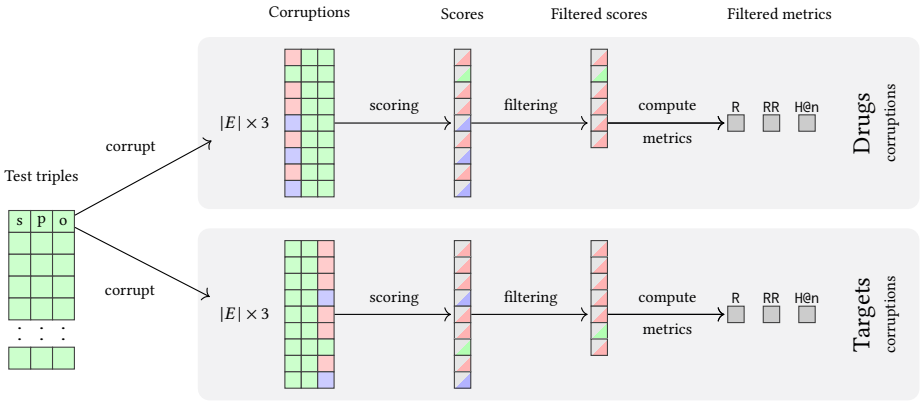

Figure 4: Evaluation protocol of KGE models for a single drug target testing instance. Note: R represents rank, RR represents reciprocal rank, and H@n represents Hits@n

Table 2: Statistics of entities, relations, facts and drug-target fact count per split of KEGG50k dataset

Dataset	Entities	Relations	Facts	DT-Facts
KEGG50k-full	16201	9	63080	12004
KEGG50k-train	16201	9	57080	10769
KEGG50k-valid	16201	9	3000	585
KEGG50k-test	16201	9	3000	650

6 EXPERIMENTS

In this section we describe the setup of our experiments and the evaluation pipeline.

6.1 Dataset

In our experiments, we divide the KEGG dataset subset into training, validation and testing splits with ratios of 90%, 5% and 5% respectively. All the splits contain facts that describe all entities and relations, where the drug-target links are distributed among the three splits with the same ratio as the splits sizes. Table 2 shows statistics about the dataset and its splits in terms of number of entities, relations, facts and drug-target facts. The KEGG50k dataset can be downloaded from figshare [6].

6.2 Implementation

We use Tensorflow framework (GPU) along with Python 3.5 to perform our experiments. All experiments were executed on a Linux machine with processor Intel(R) Core(TM) i70.4790K CPU @ 4.00GHz, 32 GB RAM, and an nVidia Titan Xp GPU.

6.3 Evaluation protocol

KGE models are evaluated using a unified protocol that assesses their performance in the task of link prediction. Let X be the set of facts, *i.e.* triples, Θ_E be the embeddings of the set of all entities E, and Θ_R be the embeddings of the set of all relations R. The KGE evaluation protocol works in four steps (Fig. 4 shows a visual flow of the evaluation process steps for a single triple instance):

(1) Corruption Let $\mathbf{x} = (s, p, o) \in X$, then for each $\mathbf{x}$, it is corrupted $2|E| - 1$ times by replacing its subject and object entities with all the other entities in E. The corrupted triples can be defined as: $\mathbf{x}_{\text{corr}} = \bigcup_{s' \in E}(s', p, o) \cup \bigcup_{o' \in E}(s, p, o')$, where $s' \neq s$ and $o' \neq o$. These corruptions effectively provide negative examples for the supervised training and testing processes due to the Local Closed World Assumption [23].

(2) Scoring: Both original triples and their corrupted instances are evaluated using a model-dependent scoring function. This process involves looking up embeddings of entities and relations, and computing scores depending on these embeddings using the model-dependent scoring function.

(3) Filtering: It is possible that corruptions of triples may contain positive instances that exist among training or validation triples. This problem is alleviated by filtering out scores of positive instances in the triple corruptions.

(4) Computing metrics: Each triple and its corresponding subject and object corruption triples produce two sets of filtered scores following previous evaluation steps. Then, for each set of filtered scores, the KGE model computes rank, reciprocal rank, and hits@n metrics.

6.4 Experimental Setup

In the experiments, we use state-of-the-art KGE models the TransE [2], the DistMult [44], and the ComplEx [38] models compared to our customised version of the ComplEx model to perform link prediction over KEGG50k dataset in two settings. First, the general link prediction setting, where the objective is to learn rank all the link in the testing set according to their factuality compared to their all other possible corrupted assertions. Second, the drug target link prediction setting, where the same previous procedure is applied only to drug target links.

[6] KEGG50k dataset is found at: https://figshare.com/s/bbfc7b82d17e0b8b6a43

Table 3: Link prediction results over KEGG50k dataset on both general links and drug target links only. For all the metrics except for mean rank (Rank), the higher the value the better.

Model	General dataset links					Drug target links				
	Rank	MRR	Hits@1	Hits@3	Hits@10	Rank	MRR	Hits@1	Hits@3	Hits@10
TransE [2]	**192**	0.46	0.38	0.50	0.63	**81**	0.75	0.69	0.79	0.86
DistMult [44]	430	0.37	0.27	0.42	0.57	186	0.61	0.50	0.69	0.81
ComplEx [38]	506	0.39	0.31	0.43	0.57	208	0.68	0.61	0.71	0.82
ComplEx-SE (This work)	534	**0.52**	**0.45**	**0.56**	**0.68**	145	**0.78**	**0.73**	**0.81**	**0.88**

We run all the models over the previously mentioned benchmarking dataset KEGG50k. A grid search is performed to obtain best hyper parameters for each model, where the set of investigated parameters are: embeddings size $K \in \{50, 100, 150, 200\}$, margin $\lambda \in \{1, 2, 3, 4, 5\}$ for the TransE and the DistMult models, and number of negative samples $n \in \{2, 4, 6, 10\}$. All embeddings vectors of our models are initialised using the uniform Xavier random initializer [10]. For all the experiments, we use batches of size 5000, with a maximum of 1000 training iterations *i.e.* epochs. The gradient update procedure is performed using the AMSGrad optimiser [28] with a fixed 0.01 learning rate.

7 RESULTS AND DISCUSSION

Table 3 shows the output results of our experiments, where the experiments are executed in two configurations: general link prediction over all association types and link prediction over drug targets associations only.

The results show that our approach, Complex-SE, outperforms other state-of-the-art models in terms of MRR, Hits@1, Hits@3 and Hits@10 on both experiment configurations, where it achieve a better MRR with a 6% margin in predicting general links and a 3% MRR margin over other models in predicting drug-target links.

The link prediction task is executed such that for each investigated possible drug-target association such as (*Aspirin*, *Drug-Target*, *COX2*) each model is required to answer two questions: (1) Which drug targets *COX2*? and (2) Which target does the drug *Aspirin* target? A model has to choose an answer from the set of all vocabulary entities, where all correct answers except for drug *Aspirin* and target *COX2* are removed. The answers to these questions are formatted as a rank, where the model is required to position the correct answer in the first place in its rank to achieve perfect accuracy as shown in Fig. 4, which presents the flow of the link prediction evaluation pipeline for one test instance. In this setting, a random baseline model would choose the right answers in the first position of the rank with a probability of $\frac{1}{|E|}$, where $|E|$ is the size of the set of all entities vocabulary, this is equal to 16201 in our experiments.

Our knowledge graph embedding model, ComplEx-SE, is able to identify the correct answers for both of the previous questions with a mean reciprocal rank of 0.78, where it identifies the correct answer within the rank with probabilities of 0.73, 0.81, 0.88 at the first, the third, the tenth positions respectively.

The use of knowledge graph embedding approaches such as ComplEx-SE enables predicting new associations for both new drugs and new targets since it does not depend on their interaction profiles. KGE models are also capable of learning different types of associations between entities of different types with no extra configurations as shown in the general link prediction configuration in table 3. For example, they can be used to identify the relation between proteins and pathways, or the relation between drugs and pathways. This can lead to discovering further unknown activities for both drugs and proteins with no extra computational cost.

8 ACKNOWLEDGEMENTS

This work has been supported by Insight Centre for Data Analytics at National University of Ireland Galway, Ireland (supported by the Science Foundation Ireland grant 12/RC/2289). The GPU card used in our experiments is granted to us by the Nvidia GPU Grant Program.

9 CONCLUSIONS AND FUTURE WORK

In this work, we introduced the use of knowledge graph embedding models for predicting drug targets using currently available drug knowledge bases, where we formulated the problem as a link prediction task over drug targets centred knowledge graphs. We have created a knowledge graph dataset, KEGG50k, from KEGG database, which is centred around drugs and their targeted genes, disease and reaction pathways. We then used this dataset to evaluate the predictive accuracy of knowledge graph embedding models.

We proposed a knowledge graph embedding approach, ComplEx-SE that is a customised version of the ComplEx model with a square error based loss function, and we showed by empirical evaluation that our approach outperforms other state-of-the-art knowledge embedding models in the task of predicting drug target links over KEGG database. Our results showed that the ComplEx-SE approach is able to identify drug target link with a mean reciprocal rank of 0.78 with a 3% margin better than other state-of-the-art knowledge graph embedding models. Results also showed that our approach is able to provide a rank of 16201 possible drug-target association statements with only one true statement, where it identifies the true state with probabilities of 0.73, 0.81 and 0.88 at the first, the third and the tenth positions of the rank.

Despite the growing body of research on computer based approaches for predicting drug targets, our objective in this work was limited to evaluating knowledge graph embedding approaches and identifying their optimal techniques for predicting drug-targets. However, in future work, we aim to perform a comparison between our knowledge graph embedding technique and other state-of-the-art drug target discovery computer based approaches based on the benchmarking dataset and evaluation metrics. We also aim to perform in-lab experimental evaluation for the top predicted drug target associations that is not in the currently public available knowledge bases to validate possible new undiscovered drug targets.

REFERENCES

[1] Ted T Ashburn and Karl B Thor. 2004. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery* 3, 8 (2004), 673.

[2] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *NIPS*. 2787–2795.

[3] Joanne Bowes, Andrew J Brown, Jacques Hamon, Wolfgang Jarolimek, Arun Sridhar, Gareth Waldron, and Steven Whitebread. 2012. Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nature reviews Drug discovery* 11, 12 (2012), 909.

[4] Anne Corbett, James Pickett, Alistair Burns, Jonathan Corcoran, Stephen B Dunnett, Paul Edison, Jim J Hagan, Clive Holmes, Emma Jones, Cornelius Katona, et al. 2012. Drug repositioning for Alzheimer's disease. *Nature Reviews Drug Discovery* 11, 11 (2012), 833.

[5] Michael Dickson and Jean Paul Gagnon. 2009. The cost of new drug discovery and development. *Discovery medicine* 4, 22 (2009), 172–179.

[6] Jürgen Drews. 2000. Drug Discovery: A Historical Perspective. *Science* 287, 5460 (2000), 1960–1964. https://doi.org/10.1126/science.287.5460.1960

[7] Michel Dumontier, Alison Callahan, Jose Cruz-Toledo, Peter Ansell, Vincent Emonet, François Belleau, and Arnaud Droit. 2014. Bio2RDF Release 3: A larger, more connected network of Linked Data for the Life Sciences. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.* 401–404.

[8] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31, 3 (2010), 59–79.

[9] Matt Gardner and Tom M. Mitchell. 2015. Efficient and Expressive Knowledge Base Completion Using Subgraph Feature Extraction. In *EMNLP*. The Association for Computational Linguistics, 1488–1498.

[10] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS (JMLR Proceedings)*, Vol. 9. JMLR.org, 249–256.

[11] Ming Hao, Stephen H Bryant, and Yanli Wang. 2017. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Scientific reports* 7 (2017), 40376.

[12] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* 45, D1 (2017), D353–D361.

[13] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44, D1 (2016), D457–D462.

[14] Ni Lao and William W. Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine Learning* 81, 1 (2010), 53–67.

[15] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Chris Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal* (2014).

[16] Linxin Li, Olivia C Geraghty, Ziyah Mehta, Peter M Rothwell, and Oxford Vascular Study. 2017. Age-specific risks, severity, time course, and outcome of bleeding on long-term antiplatelet treatment after vascular events: a population-based cohort study. *The Lancet* 390, 10093 (2017), 490–499.

[17] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. 2015. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 31, 12 (2015), i221–i229.

[18] Jian-Ping Mei, Chee-Keong Kwoh, Peng Yang, Xiao-Li Li, and Jie Zheng. 2012. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29, 2 (2012), 238–245.

[19] George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[20] Sameh K. Mohamed, Vít Novácek, and Pierre-Yves Vandenbussche. 2018. Knowledge base completion using distinct subgraph paths. In *SAC*. ACM, 1992–1999.

[21] Emir Muñoz, Vít Novácek, and Pierre-Yves Vandenbussche. 2016. Using Drug Similarities for Discovery of Possible Adverse Reactions. In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*. AMIA. http://knowledge.amia.org/amia-63300-1.3360278/t004-1.3364525/f004-1.3364526/2499657-1.3364713/2500122-1.3364708

[22] André CA Nascimento, Ricardo BC Prudêncio, and Ivan G Costa. 2016. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics* 17, 1 (2016), 46.

[23] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2016), 11–33.

[24] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. 2016. Holographic Embeddings of Knowledge Graphs. In *AAAI*. AAAI Press, 1955–1961.

[25] Rawan S Olayan, Haitham Ashoor, and Vladimir B Bajic. 2017. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 34, 7 (2017), 1164–1173.

[26] Richard Qian. 2013. Understand Your World with Bing. http://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/ Bing Blogs.

[27] Yan Qiao, Tingting Yang, Yong Gan, Wenzhen Li, Chao Wang, Yanhong Gong, and Zuxun Lu. 2018. Associations between aspirin use and the risk of cancers: a meta-analysis of observational studies. *BMC cancer* 18, 1 (2018), 288.

[28] Sashank Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the Convergence of Adam and Beyond. In *ICLR*.

[29] Ayeshah A Rosdah, Jessica K. Holien, Lea MD Delbridge, Gregory J Dusting, and Shiang Y Lim. 2016. Mitochondrial fission–a drug target for cytoprotection or cytodestruction? *Pharmacology research & perspectives* 4, 3 (2016), e00235.

[30] Peter M Rothwell, F Gerald R Fowkes, Jill FF Belch, Hisao Ogawa, Charles P Warlow, and Tom W Meade. 2011. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet* 377, 9759 (2011), 31–41.

[31] Peter M Rothwell, Michelle Wilson, Carl-Eric Elwin, Bo Norrving, Ale Algra, Charles P Warlow, and Tom W Meade. 2010. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *The Lancet* 376, 9754 (2010), 1741–1750.

[32] Amit Singhal. 2012. Introducing the Knowledge Graph: things, not strings. "https://googleblog.blogspot.ie/2012/05/introducing-knowledge-graph-things-not.html" Google Official Blog.

[33] Lekha Sleno and Andrew Emili. 2008. Proteomic methods for drug target discovery. *Current opinion in chemical biology* 12, 1 (2008), 46–54.

[34] Walter Sneader. 2005. *Drug discovery: a history*. John Wiley & Sons.

[35] Georg C Terstappen, Christina Schlüpen, Roberto Raggiaschi, and Giovanni Gaviraghi. 2007. Target deconvolution strategies in drug discovery. *Nature Reviews Drug Discovery* 6, 11 (2007), 891.

[36] James Hendler Tim Berners-Lee and Ora Lassila. 2001. The Semantic Web, A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American: https://www.scientificamerican.com/article/the-semantic-web/. Retrieved: 2017-04-21.

[37] Théo Trouillon and Maximilian Nickel. 2017. Complex and Holographic Embeddings of Knowledge Graphs: A Comparison. *CoRR* abs/1707.01475 (2017).

[38] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML (JMLR Workshop and Conference Proceedings)*, Vol. 48. JMLR.org, 2071–2080.

[39] Twan van Laarhoven, Sander B Nabuurs, and Elena Marchiori. 2011. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 21 (2011), 3036–3043.

[40] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* 34, suppl_1 (2006), D668–D672.

[41] Lei Xie, Li Xie, Sarah L Kinnings, and Philip E Bourne. 2012. Novel computational approaches to polypharmacology as a means to define responses to individual drugs. *Annual review of pharmacology and toxicology* 52 (2012), 361–379.

[42] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. 2008. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, 13 (2008), i232–i240.

[43] Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto. 2010. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, 12 (2010), i246–i254.

[44] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *ICLR*.